The characterisation of engineering activity through email communication and content dynamics, for support of engineering project management

Snider, C. (Corresponding), Skec, S., Gopsill, J.A., Hicks, B.J.

University of Bristol

University of Bath

United Kingdom

University of Zagreb

Croatia

Abstract

Significant challenge exists in the effective monitoring and management of engineering design and development projects. Due to traits such as contextual variation and scale, detailed understanding of engineering projects and activity are difficult to form, with monitoring hence reliant on interpretation of managerial personnel and adherence to defined performance indicators.

This paper presents a novel approach to the quantitative monitoring and analysis of engineering activity through computational topic identification and analysis of low-level communication data. Through three metrics of communication activity, this approach enables detailed detection and tracking of activity associated with specific project work areas. By application to 11,832 emails within two industry email corpora, this work identifies four distinct patterns in activity, and derives seven characteristics of communication activity within engineering design and development. Patterns identified are associated with background discussion, focussed working, and the appearance of issues, supporting detailed managerial understanding. Characteristics identified relate to through-process norms against which a manager may compare and assess.

Such project-specific information extends the ability of managers to understand the activity within their specific project scenario. Through detailed description of activity and its characteristics, in tandem with existing toolsets, a manager may be supported in their interpretation and decision-making processes.

1 Introduction

With the increasing technological capability of the modern world, the tools and processes of engineering have become increasingly digital. A lower barrier to entry, higher precision, and ease-of-use have moved design and modelling to computerised packages, while digital analysis approaches such as finite element analysis have led to lower manual effort with increased breadth and detail in results. This trend has led to process and output improvements (Cantamessa et al. 2010; Banker et al. 2006), but has also paralleled a trend towards ever increasing project complexity. Single engineering projects have potential to involve thousands of engineers, working on tens of thousands of systems and components, spread over multiple countries or continents (Watson 2012). Even in smaller projects, complexity and risk (Earl et al. 2005; Chapman & Ward 1996) have potential to cause delays, budget over-run, and quality reduction (Xia & Lee 2004); issues that are exacerbated as project scale increases (Floricel & Miller 2001). While technological growth has enhanced engineering capability, the challenges facing project management have grown.

Typically, engineering projects are managed through the *iron triangle* – cost, quality, and time – with the outputs and processes of the project judged against targets associated with each (Lavagnon 2009; Collins & Baccarini 2004). There is broad recognition, however, that such measures provide only a partial, lagging picture (Atkinson 1999; Toor & Ogunlana 2010; Schmidt et al. 2001). While forming a post-hoc measurement, it is the amalgamation of underlying factors and circumstances that lead towards such criteria as on-budget, on-time, and on-quality (Pinto & Slevin 1987; Snider, Gopsill, et al. 2015; Cooke-Davies 2002); and so it is these underlying factors that must be improved to ensure high project performance. Due to the bespoke nature of engineering projects this is a particularly challenging task (Engwall 2002); every project will succeed or fail based on different factors, and hence monitoring and management must be sensitive to the specific project context to ensure effectiveness.

This paper aims to support the management of engineering design and development projects through a method for detailed monitoring of engineering work and associated patterns in activity, extracted directly from low-level data analytics. This provides potential benefit in two streams; project planning based on post-hoc analysis of historical projects, and real-time monitoring of ongoing live data.

1.1 Generation and Use of Project-Specific Information

Much research in recent years has focused on the generation of detailed project-specific information through manual and automatic information gathering methods, enabling the application of fine-grain analytic approaches to understand in detail the activity of engineers working within engineering design and development environments. This mirrors and leverages the advantages enabled by digital capture and analysis of high volume, velocity, and variety data seen in other fields, where extensive and broad-spectrum data capture and automatic analysis have enabled a new paradigm in analytic capability (Eagle & Pentland 2006; Labrinidis & Jagadish 2012). Aligned work within the field of design research has included, for example, manual work sampling through direct engineer input (Robinson 2010) and subsequent monitoring and prediction of performance through application of network analytics and alignment with survey results (Škec et al. 2017), email and digital work monitoring to determine team interaction through network analysis (Uflacker & Zeier 2011), and monitoring of information flow and project classification through analyses of communication networks associated with process stage (Parraguez et al. 2015).

Such information gathering and detailed analyses lie in contrast to many methods employed in engineering design research. Utilising such methods as logbook coding (Snider, Mcalpine, et al. 2015; McAlpine et al. 2009), retrospective interviews and questionnaires (Cross & Cross 1998; Ahmed et al. 2003), ethnographic observation (Hales 1987; Wallace & Ahmed 2003), and protocol study (Gero &

Tang 2001; Dorst & Cross 2001), the observational activity monitoring frequently employed in design research struggles when approaching large-scale implementation and analyses or highly reactive, near real-time generation of understanding.

Through implementation of highly detailed, digital activity monitoring methods, significant benefits may be found in two streams. Firstly, in support of learnings from post-hoc analyses and study of archival engineering data, the volume, velocity, and variety of data gathered through digital analyses enable both a finer-grain in analysis output and additional learnings from new analytic capabilities (Chen et al. 2012). Such information may be used for project diagnoses and future project planning. Second, by enabling broad-spectrum automatic data analyses in the digital space the potential for highly reactive project monitoring is created (Snider, Gopsill, et al. 2015), in which live analyses provide engineering managers with an evidence base for their decision making processes, while reducing investigative effort.

1.2 Aim and Significance of the Work

Representing the communication network and communication activity employed within a project, the study of email transaction has frequently been utilised within research as a medium for study of worker activity, and consequent project performance. Indeed, the structure of communication networks and patterns of transaction within have validated alignment with process stage (Parraguez et al. 2015), project output characteristics (Dewar & Dutton 1986), and project performance (Aral et al. 2007; Rodan & Galunic 2004; Landaeta 2008; Patrashkova-Volzdoska et al. 2003).

This work aligns with this thinking, utilising the information transmitted within engineering communication networks to imply project-specific information. Here however, where other research often applies at a transactional level through analysis of the characteristics of the communication network structure (see Uflacker & Zeier 2011; Parraguez et al. 2015; Gruhl et al. 2004; Tang et al. 2010; Aral et al. 2007), this work aims to monitor the activity of engineers through the combinatory study of the transactional characteristics of email (i.e. sender / receiver / time / cc) with their *content.* In this manner it aims to study association between the way in which certain content - that associated with specific *work areas*, termed *topics* - and discursive activity are related.

Based on the knowledge that the role of specific information topics in the project context may influence their manner of communication within the network (von Hippel 1998; Romero et al. 2011; Cha et al. 2009), this work performs an exploratory analysis and classification of the communication activity patterns apparent in discussion within two engineering design and development projects. Through this analysis, it aims to identify and associate types of activity pattern with types of content in engineering work, and with engineering project and process characteristics. In particular, it attempts to identify commonalities in patterns across differing engineering contexts, and thus produce results with broader generalizability.

The benefit of identification of activity patterns associated with individual work areas and topics lies in increased understanding of the role and contextual progress of activity within a specific project. For example, dependent on level in system hierarchy, emergence of issues, process stage, or whether activity is conforming to “normality”, different communication patterns may be expected to appear. Identification of such patterns and their implication therefore provides scope to increase understanding of engineering activity through automatic analyses.

2 Analysis of Email Communication

In attempt to study activity through automatic and non-intrusive means, this work analyses the email communication sent throughout projects. Forming a key communication method within engineering (J Wasiak et al. 2010; Gupta et al. 2009) and 14% of engineer work in itself (Robinson 2012), email is a proven route to understanding social interaction, content formation, and user effectiveness (O'Kane 2007). As such, emails provide a strong representation of underlying activity – through the study of the transmission and content of email communication, there is potential to understand the engineering activity by which it was created. For this reason it is highly studied in literature, addressing such areas as topic identification (Allan 2002), information diffusion amongst project members (Aral et al. 2007; Iribarren & Moro 2009; Wu et al. 2004), context of content such as sentiment (Byron 2008; Pang & Lee 2008) and authority (Jones et al. 2014), identification of communities and social groups (Johansen et al. 2007), and direct monitoring of project performance through email terminology use (Munson et al. 2014).

Within engineering design, email has been studied to clarify the time commitment it demands (Robinson 2010; Robinson 2012) or the purpose of emails sent (James Wasiak et al. 2010), but has not focussed on the relationship between individual topics and the activity that accompanies their discussion. In addition, there has been little research studying such with an industrial focus, particularly within the engineering domain.

To enable this analysis this work performs a number of steps. First, there is a need to identify relevant topics within the specific datasets under study. Second, the occurrence of each topic must be tracked through each project. This is here performed using three metrics, see Table 2, each with a different implication for the project manager. Next, it uses outputs of this analysis to identify patterns of two forms – dynamic traces in the discussion of individual work areas and topics, and characteristics in the appearance of these traces through the engineering process. The implication or meaning of these patterns in the project context is then identified and validated through study of specific project content and comparison between the two projects under study. The analysis process for all data is as described in Table 1 and detailed in the following sections.

Table 1: Analysis Process

	Dataset A	**Dataset B**
1	Data extracted, cleaned, and processed for analysis; see Section 2.1	
2	Initial topics extracted using *tf-cidf* algorithm; see Section 2.2	
3	Extracted topics manually parsed; see Section 2.2, Section 5	
4	Analysis metrics algorithmically implemented; see Section 2.3 – 2.5	
5	Activity patterns and stage characteristics identified; see Section 3, Section 4	
6		Validation of pattern and characteristic appearance in second dataset; see Section 3, Section 4
7	Content of communications relating to each pattern studied to determine project implication; see Section 3 – 5	

The work presented here builds on previous publications, which placed their focus on technique development and feasibility (Snider et al. 2016; Jones et al. 2015). While not used for identical purpose as in this paper, these metrics have many parallels in literature (see Gruhl et al. 2004; Guille & Hacid 2012; Romero et al. 2011; Matsubara et al. 2012; Pastor-Satorras & Vespignani 2001).

Table 2: Metrics of topic discussion and activity

Metric	Description	Interpretation
Cumulative Occurrence	Number of communications about a topic.	The proportion of communications that refer to each topic during each time period.
Relative Occurrence	Ratio of current number of topic communications to previous mean number of topic communications.	The extent above or below expected levels of discussion, based on past usage.
Occurrence Duration	Proportion of project process with high levels of communication about a topic.	The extent to which different topics are under active discussion throughout the project process.

2.1 Datasets

This work utilises two datasets; email corpora belonging to single, long-term industrial projects, allowing direct comparison between two industry contexts. As this work aims to identify consistent patterns in activity, it is vital that cross-context comparison occur. Due to the high variation in scale and branch of the engineering industry, any patterns found to be consistent between datasets demonstrate generalisability of findings and the analytical approaches applied.

Comparatively limited size and dynamism in the communication networks of industry, due to for example changes in input from workers as the process progresses, demand consideration of the propensity for the networks under study to change (Tang et al. 2010; Carley 2003). As such, the *active networks* for each company must be considered, including only those members currently eligible to send or receive a topic, and hence mirroring *susceptible-infected-recovered* models of information diffusion (Holme & Saramaki 2012; Wu et al. 2004; May & Lloyd 2001). Active network size was determined through monitoring those members with participation within the previous 4 time steps; a period of inactivity longer than this indicated a high likelihood that the member would not reappear. Summary data for each is given in Table 3.

Dataset A was taken from a large marine engineering company, relating to a long-term, large-scale engineering systems design, development, and implementation project. All members were required by policy to send all project-related emails to a project inbox, which was extracted for analysis. Emails were gathered over all stages of the process from initiation to roll-out, including 10,277 emails over 135 weeks (mean 10.8/day) between 675 involved, globally distributed members. The active network consisted of a mean of 74 members (*range 6 – 134*), with all workers having potential to input to multiple projects simultaneously.

Dataset B was significantly smaller, consisting of 1,546 emails from a small software development company, and relating to a single software development project over a two-year period. All employees were required to archive emails in project-specific inboxes, which were copied and collated. Emails were sent or received by 78 persons, of which 6 were co-located core employees of the company within a single office, and others were nationally distributed. The frequency of emails was correspondingly low (1.98/day), but reached a mean of 10.8/day during busiest times. Employees were not working solely on this project, as indicated by bursts and stalls in email frequency (Figure 1). The active network consisted of a mean of 13 active members (*range 2 – 28*).

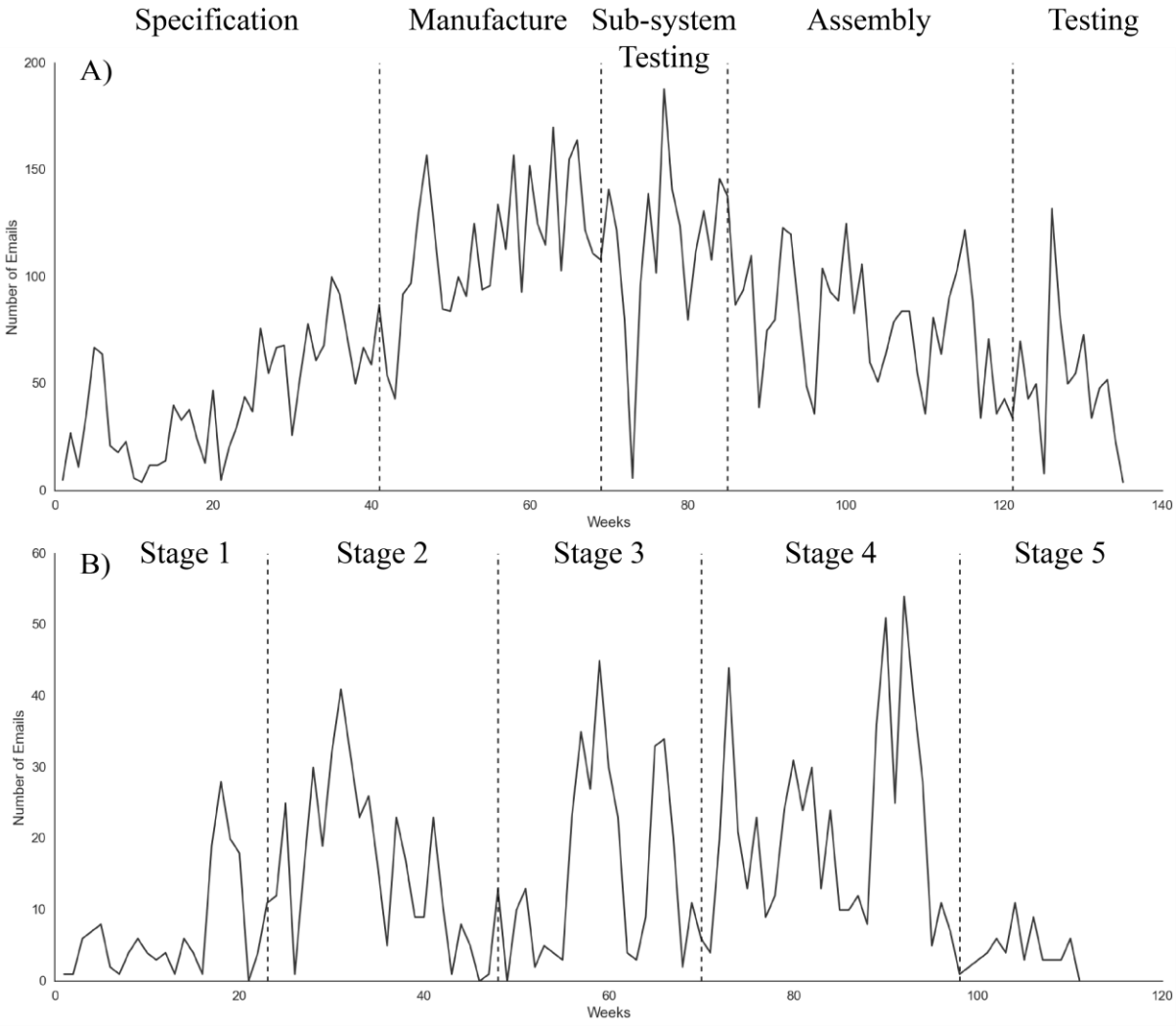


Figure 1: Email frequency with time; (above) Dataset A, (below) Dataset B

Discrete stages within Dataset A were determined through interview with project members, allowing the project to be separated into specification, manufacture, sub-system testing, assembly, and testing (see Figure 1A). Dataset B followed an agile methodology, as are common in software development environments (Highsmith & Cockburn 2001; Fernandez & Fernandez 2008), with periods of work occurring in line with deadlines for deliverables and calls for development. As such it cannot be discretised into individual stages, with boundaries instead placed during the periods of lower activity (see Figure 1B), which typically follow deadlines for deliverables and during which workers focused on projects other than that studied. Note that as stage boundaries are often fluid in nature, a one week overlap was assumed between each in all analysis.

Both Datasets demonstrated correlation between active network size and frequency of emails sent (*Dataset A, $r = 0.734$, $p < 0.0001$; Dataset B, $r = 0.644$, $p < 0.0001$*), suggesting a consistency in per-engineer email frequency. Based on active network size, per-person, per-week email frequency was also consistent (*Dataset A, 1.02 email per-person per-day; Dataset B. 1.06 emails per-person per-day*), although these values assume an equal role of all active members each week, hence ignoring that many may lie on the periphery of project activity. While the quantity of email to be expected in a given project is highly dependent on the project context, such consistency suggests a level of normality between the two contexts in their email usage.

While company mandate dictated the storing of emails within the extracted mailboxes for each project, there is potential for workers to discuss multiple projects in a single email, hence excluding an unknown quantity of communications. Further, an unknown quantity of communication likely occurred through face-to-face or other means, as is typical in engineering industry (Robinson 2012). Quantification and capture of such further communications, through extension of the dataset, would provide further confidence and detail in the results of this work.

Table 3: Dataset Summary

Process Stage	**Number of Emails**		**Period (Weeks)**		**Emails per Week**	
Dataset A / B	A	B	A	B	A	B
Whole Process	10,277	1,546	136	112	76.1	13.9
Specification / Stage 1	1,642	148	40	21	41.1	6.73
Manufacture / Stage 2	3,230	400	28	25	111	14.8
Sub-system Testing / Stage 3	1,935	350	16	22	114	14.6
Assembly / Stage 4	3,027	608	36	28	81.8	20.3
Testing / Stage 5	798	76	15	16	53.2	5.07

2.2 Topic Identification

Analysis within this work is based on individual *topics* extracted from email content, where a topic refers to a single work area or subject under discussion. As topics have potential to be highly project-specific, this must be performed on a case-by-case basis.

Describing the specific information under discussion, the topics tracked by each metric were denoted through the appearance of specific phrases within the emails of each dataset. Numerous methods for topic identification exist (Allan 2002; Cataldi et al. 2010; Coursey & Mihalcea 2009; AlSumait et al. 2008), with this work utilising *term frequency – cumulative inverse document frequency (tf-cidf)* (Gruhl et al. 2004), an algorithm similar to the widely employed *term frequency – inverse document frequency (tf-idf)* method (Spärck Jones 1972; Robertson 2004), here selected due to its ability to identify key topics within individual project timespans, and its identification of topics of varying generality. While this method omits factors such as context and semantic similarity (Wang et al. 2014; Brants et al. 2002; Kuhn et al. 2007), it excels in breadth of topics extracted, thereby allowing higher numbers of topics to be generated, a breadth of types of pattern in activity to be identified, and hence aligning with the core purpose of this paper. Further, topics assigned as sematic groupings by algorithms such as latent semantic analysis (Dumais et al. 1988; Dumais 2004), although providing greater detail in the similarity and subject matter under discussion, have propensity to change in size, membership, and shape through the project timeline, thus making their consistent and robust tracking a significant challenge. While semantic topic modelling through are considered valuable further work, term extraction through *tf-cidf* meets the topic identification goals of this exploratory classification, and mirrors topic and information diffusion modelling methods employed in other fields (Gruhl et al. 2004).

Formally, where *tfcidf(t)* is document score in a given time period *t*, *n* is total number of time periods, and $F_T(t)$ is frequency of occurrence of a topic in period *t*:

$$tfcidf(t) = \frac{(n-1)F_T(t)}{\sum_{t=0}^{n-1} F_T(t)}$$

To generate the topic list used for analysis, the *Tf-cidf* algorithm was applied to each dataset. A minimum time period, *t*, of one week was used to ensure sufficient email quantities for analysis. Following values used in other research (Gruhl et al. 2004), one-gram (single-word) topics were identified using thresholds of $tfcidf(t) \geq 12$ and $tf \geq 3$, and bi-gram and tri-gram (two- and three-word) topics were identified from thresholds of $tfcidf(t) \geq 10$ and $tf \geq 3$.

Emails from each mailbox were algorithmically parsed to remove metadata, duplicates, and stop-words, and to extract subject line, content including signature, sender / receivers, and date sent. While not directly relevant as discussed content email signatures typically contain company information, allowing monitoring of activity associated with companies and individual projects as extracted by the *Tf-cidf* algorithm.

Potential topics highlighted by the *Tf-cidf* algorithm were manually parsed by an engineer to produce a single topic word list, where each topic referred to a system, role, person, object, or concept within the project process, output, personnel, or management. Other terminology, such as words relating to the working lexicon of the engineers, were removed. This manual process is highly dependent on the parser, and while effort was made to be inclusive, in application to industry it is vital that those topics monitored are carefully selected from the *Tf-cidf* output by personnel with a breadth of experience across the sectors each project may span. Here, effort was made to include topics of varying type to encourage broad classification. Parsing produced topics as per Table 4, where an individual topic list was generated for each dataset.

Table 4: Topic Identification Results

	One-gram		**Bi-gram**		**Tri-gram**		**Total**	
	A	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Unique	26,010	8,796	66,097	14,099	27,510	4,701	119,617	27,596
Candidate topics	1275	906	465	278	85	70	1825	1254
Selected	63	38	216	71	55	20	334	129
Example topics (Dataset A)	Company 1B Outfitting Pressure Spares Converter Voltage Engine		Electrical Systems Propulsion motor Propulsion transformer Cold Ironing Project implementation Interface list Purchasing department		Gas combustion unit Lotus notes release Onboard test procedure Spare part list Risk assessment form Electric design section Network switch boxes			
Example topics (Dataset B)	Company 2A1 Invoice Schedule Server Workshop		Data setup System 2A setup Project costs Generic access Human Resources		Long haul flight System 2A API setup Suitable test data Working day length			
Example terms removed	Discuss, accept, kind, submitted, recommend, meet		Meeting held, information contained, good afternoon		Dates, phone numbers, personnel names, *"points raised wrt"*			

2.3 Analysis Metric – Cumulative Occurrence

The proportion of communications in a given time period that refer to a given topic give an indication of its accompanying level of activity. Those topics that demand higher proportions of communications suggest a higher focus on that particular area. Although mention in an email does not necessarily indicate that substantial work has occurred, it does indicate that the topic has been the subject of attention for multiple engineers.

Cumulative occurrence is defined as the following, where $F(t)$ is normalised cumulative occurrence of a topic in time period t, $F_e(t)$ is number of emails in time period t, and $F_T(t)$ is number of emails containing a topic in time period t:

$$F(t) = \frac{F_T(t)}{F_e(t)}$$

Cumulative occurrence is used in this work to distinguish between topics that frequently receive attention throughout the project and those that are more rarely or briefly discussed. Topics have a slight tendency to lower values. The highest and lowest cumulative occurrence topics are given in Table 5. Whilst highest scoring topics are typically those that appear within email signatures, such as company and project names, a number of more specific topics also have higher cumulative occurrence, thus indicating work areas which typically receive higher attention.

Table 5: High and low cumulative occurrence topics, by mean value over topic life.

High Cumulative Occurrence (% of emails in time period)		**Low Cumulative Occurrence (% of emails in time period)**	
Topic	**Mean**	**Topic**	**Mean**
Dataset A			
Company Acronym 1C	68.3	Breaker open	0.388
Company Acronym 1B	51.4	Call system	0.442
Project 1A	50.9	Gas heaters	0.506
Project Acronym 1A	50.4	Rated load	0.517
Company 1C	50.1	Order amendment	0.564
Other high scoring topics: Offshore division, specification, requirement, electric section, project manager, spec, EPS, converter FAT schedule			
Dataset B			
Company Acronym 2B	53.6	Supplier link	1.54
Company Acronym 2A	44.3	Occupancy rules	2.15
Company 2A1	39.7	Rooming lists	2.58
Company 2C	32.7	Security	3.28
Company 2A2	32.5	Webservices team	3.30
Other high scoring topics: systems workshop, fare query, return journey, specific fare, test results			

2.4 Analysis Metric – Relative Occurrence

All topics will have an *expected* occurrence for a given project with, for example, company names expected to appear more frequently than specific low-level systems (see Table 5). Deviation from this expected value may be of particular use to a manager, potentially being indicative of issues surrounding the topic, a lack of due attention, a change in general work focus, or a change in topics forming the core of the project.

As the *expected* occurrence level of a topic cannot easily be predicted, an estimate for a given point in time is formed through the average frequency of topic appearance over a certain prior time period, to which current usage is compared. This ratio is termed *relative occurrence* and is formally defined as below, where O_T is relative occurrence of a topic, $F_T(t)$ is topic frequency in time period t, $F_e(t)$ is frequency of all emails, n is the current time period, s is a short-term threshold describing current discussion, and l is a long-term threshold used to generate the estimate of expected discussion.

$$O_T = \frac{\sum_{t=n-s}^{n} F_T(t)}{\sum_{t=n-s}^{n} F_e(t)} \cdot \frac{\sum_{t=n-l}^{n} F_e(t)}{\sum_{t=n-l}^{n} F_T(t)}$$

A topic is of higher-than-expected occurrence when the shorter-term proportion, s, is higher than the longer-term, l, or $O_T > 1$. As projects are of varying length and pace values for s and l are case-dependent; for the longer-term projects studied here, a value of $s = 2$ weeks has been found to be representative of recent work. The value for l should be chosen to provide useful data within the specific case. When too small relative to the value of s, $O_T > 1$ frequently, typically for every rise in topic frequency that occurs (24 cases found in Fig.2A). This has the potential effect of highlighting small,

background bursts of activity as a significant event. Conversely, a large *l* in comparison to *s* tends to a value of *l* at the topic mean. For a topic with higher variance in occurrence through the project process, such as in Figure 2C, this provides little sensitivity at the higher and lower ends of the topic range and has potential to omit meaningful events. Here, values of $s = 2$ and $l = 8$ were found to be of appropriate sensitivity (see Figure 2B).

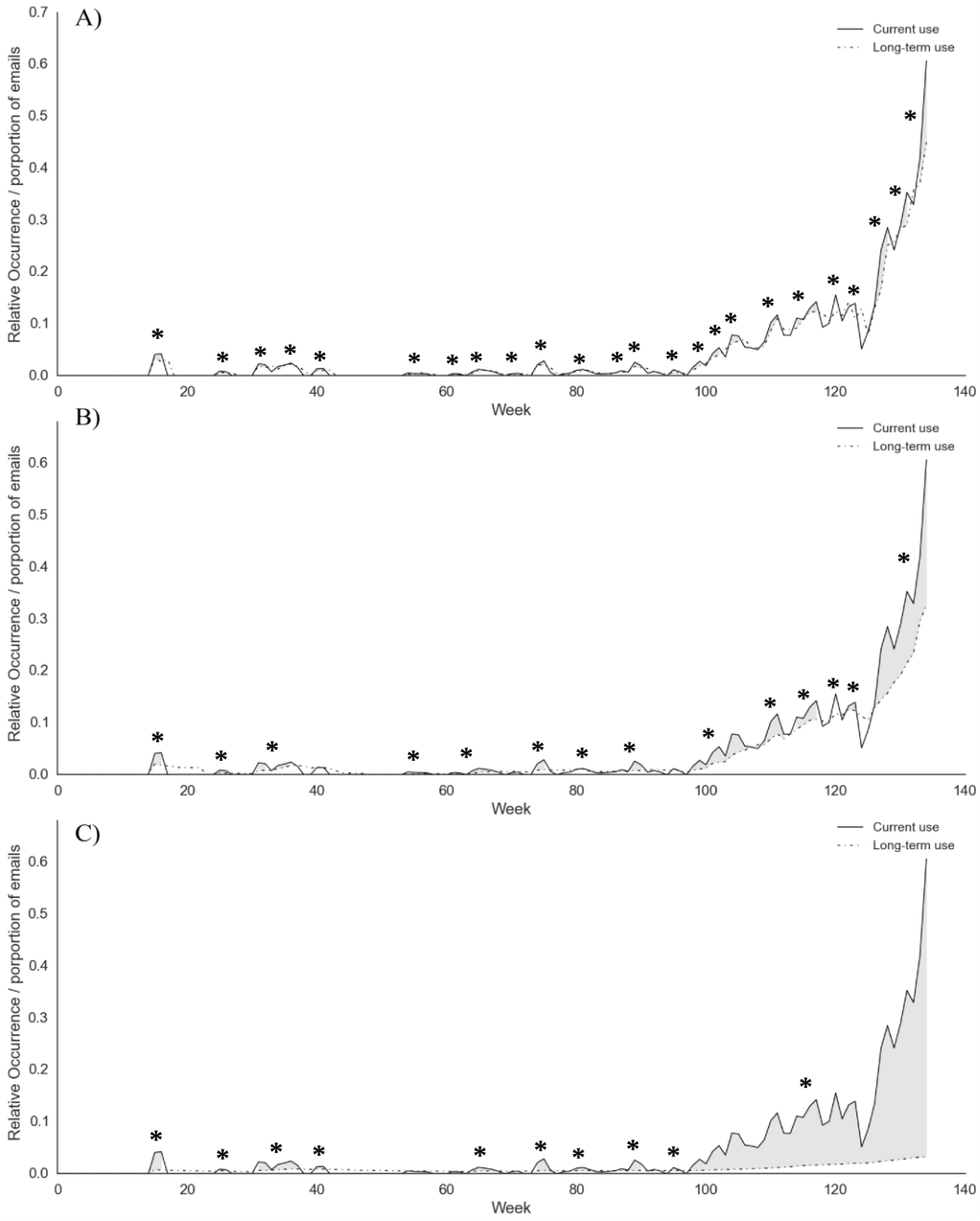


Figure 2: Effect of change in long-term threshold for topic *"project implementation"* with short-term threshold, $s = 2$. Shaded areas indicate a value of $O(T) > 1$; "*" indicates a period in which $O_T > 1$. A) 3 week length; B) 8 week length; C) Whole project length.

2.5 Analysis Metric – Occurrence Duration

Each individual topic will have a varying lifespan through the project, during which it is actively discussed by workers. By identifying typical length of active discussion, where $O_T > 1$, and relation to patterns in discussion and project stage, project norms may be identified.

By identifying the periods through which $O_T > 1$, for a topic, this work identifies the lifespan of a topic over which active discussion is occurring, as opposed to periods in which topics are mentioned without significant associated worker effort. This is termed *Occurrence Duration*, P_T, and is formally defined by the following, where $O_T(t)$ is relative occurrence in time period t, f is the first time period of occurrence, and n is number of time periods.

$$P_T = \frac{\sum_{t=f}^{n} f(O_T(t))}{n} \quad ; \quad where \quad f(O_T(t)) = \begin{cases} 1, & O_T(t) \geq 1 \\ 0, & O_T(t) < 1 \end{cases}$$

A topic with higher occurrence duration is actively discussed for a longer proportion of the project life, while a shorter duration indicates that activity is more specific to individual time periods. Duration for most is generally low. As duration increases so do the generality of the topics, describing larger components, systems, and the names of involved companies (see Table 6).

Table 6: Highest and lowest occurrence duration scores for each dataset

High Occurrence Duration		**Low Occurrence Duration**	
Topic	**Occurrence Duration**	**Topic**	**Occurrence Duration**
Dataset A			
Company Acronym 1C	55.3	PIDS	1.54
Project 1A	52.3	Electric section	1.54
Company Acronym 1B	51.5	Planning dept	1.54
Pressure	50.8	Telecoms	2.31
Project Acronym 1A	50.0	Bypass switch	2.31
Other high scoring topics: voltage, EPS, offshore division, spec, propulsion motor, specification			
Dataset B			
Company 2A	52.3	Suitable test data	1.80
Company Acronym 2B	52.3	Systems workshop	2.70
Accuracy	52.3	Daily working sessions	2.70
Company 2B	51.4	PAX short haul	2.70
Company 2A	48.6	API setup	3.60
Other high scoring topics: accuracy, schedule, training, booking, flights, pricing, invoice, staff			

3 Patterns in Topic Activity

Through analysis of these metrics in each dataset, this section aims to identify common patterns in email communication around individual topics throughout the project process. Such characterisation of consistent patterns may allow monitoring and analysis of in-progress communication, and detailed understanding of archival communication data for purposes of future planning. The following sections identify a number of characterisations, with Section 3 presenting common trends of activity on a per-topic basis, and Section 4 presenting characteristics of topic occurrence through the project life.

3.1 Background Chatter Pattern

For any ongoing project there is likely a certain quantity of background chatter – topics that occur at a consistent rate through longer periods. Such topics can be found in each project, and are characterised by a through-life median and mean value of $O_T \sim 1$ for the topic and a narrow inter-quartile range (IQR) (ie. current topic occurrence is typically close to longer-term topic occurrence). Example topics falling within this category are shown in Figure 3.

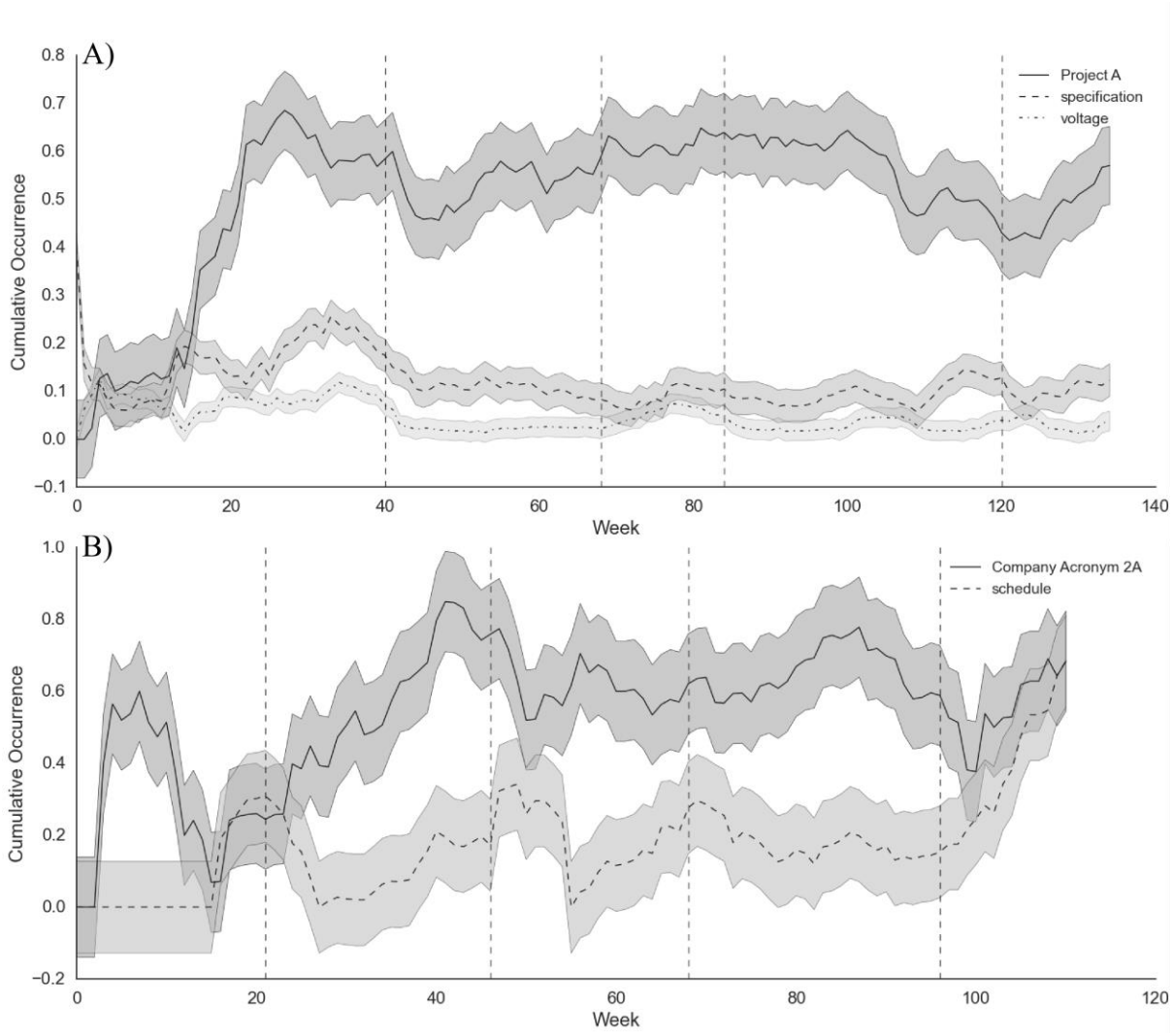


Figure 3: Background topic areas from Dataset A (above) and Dataset B (below). Line indicates longer-term occurrence. Shaded area indicates standard deviation of shorter-term occurrence.

From the data, background chatter appears either as discussion of higher-level topics (i.e. *"Project A"*, Fig. 3A) or of lower-level topics that describe systems, sub-systems, or lower-level concepts (*"specification", "voltage"*, Fig. 3A). Such patterns can also be identified in Dataset B, see Fig. 3B, although with higher variability due to smaller email frequencies in each time period. In all cases, background discussion topics are those that are pervasive through the project either as a general descriptor (i.e. *"specification"*), or as a context-specific core project area (i.e. *"voltage"*).

3.2 Single Spike Occurrence Pattern

In contrast to background chatter, some communication activity around topics occurs primarily in a single burst, here termed a *spike*. These topics have a high median and mean value of O_T due to high occurrence over a short life, and vary in amplitude dependent on cumulative occurrence, F_T, with example topics shown in Figure 4.

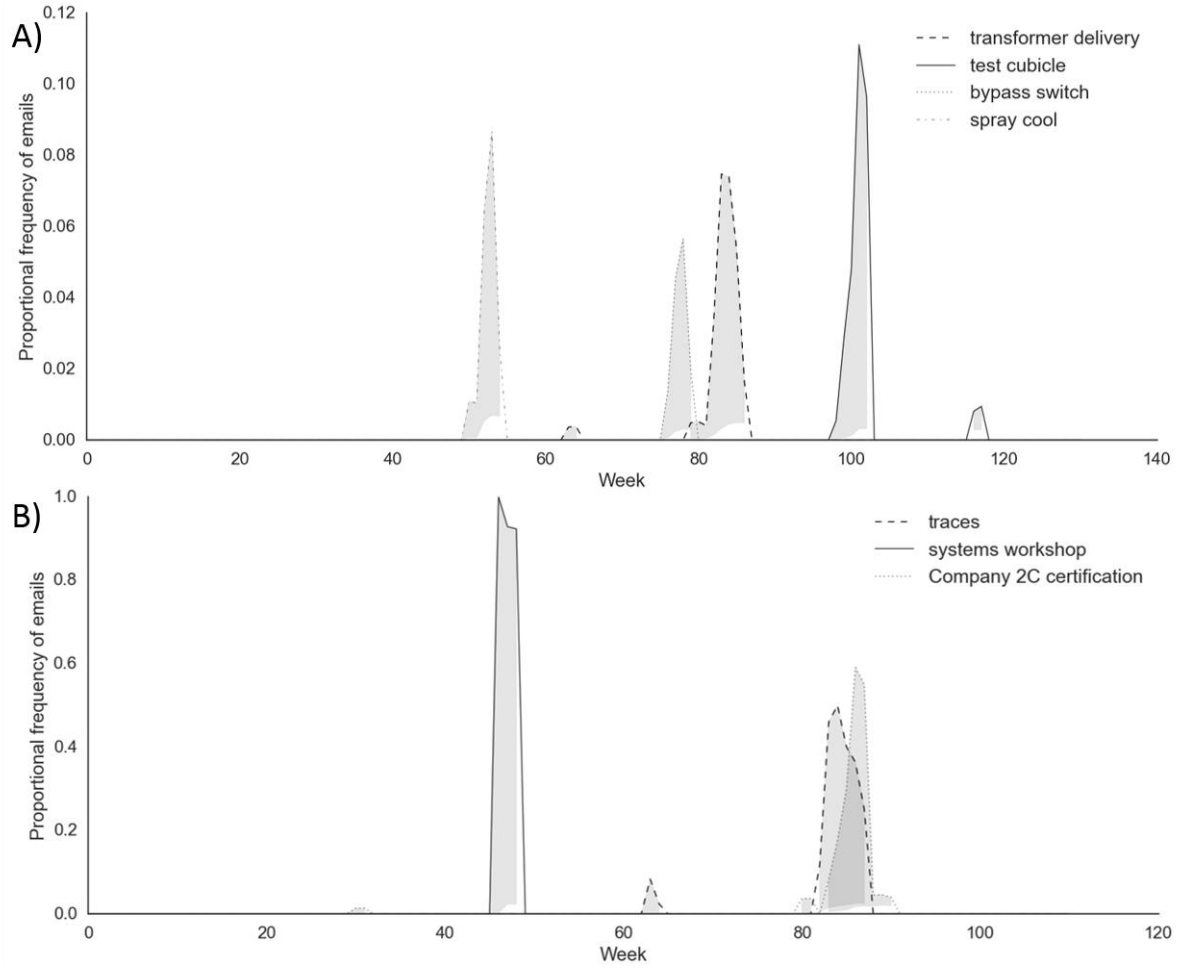


Figure 4: Single spike topic areas from Dataset A (above) and Dataset B (below). Shaded areas indicate a value of $O(T) > 1$.

A spike in topic activity represents high relative occurrence for a very brief time, with associated topics typically representing more specific topic areas, such as individual components, analyses, and lower-level systems. For a notable spike to occur, there is a need for a significant increase in cumulative occurrence of a topic in a certain time period, implying particular focus on the topic of the project workers. This may occur as part of the normal working process, where a spike simply represents the amount of communication activity required to complete the tasks at hand. For example, *test cubicle* (Figure 4A) refers to the delivery of a test platform for a low level sub-system, and is discussed only through its delivery process.

In contrast, the appearance of a spike may also represent the emergence of issues around the subject matter of the topic, in which more extensive information requests or broader input from a number of personnel are required. This interpretation can be found in literature (Gruhl et al. 2004; Myers et al. 2012), and in this case is supported by the email content studied in each project, see quotes given in Table 7. While the impact and seriousness of such events is varied, their highlighting through spike detection could provide a valuable monitoring tool for the manager.

Table 7: Email context of single spike topics

Topic	Context / Quote
Project 1 *Bypass Switch*	A misunderstanding relating to contents of requirements and work to be done. *"I have asked our engineers to review the contract requirements on us to try and understand how we are in the position"* *"From our discussion this morning this is not acceptable to [COMPANY ACRONYM 1B]*
Project 2 *Systems Workshop*	A session to directly address issues occurring within the project *"Participation in the workshop will be a significant additional unbudgeted cost for us."* *"[…] we have already incurred additional costs on this project that remain as yet unrecovered."* *"The workshop would basically be to go over all the outstanding project issues and agree a way forwards."*
Project 2 *Traces*	An issue with results of an implementation and interface with an external database, requiring discussion between supplier and user to resolve *"We'll retest here and advise"* *"Here is an update on your Master Pricer Error Message"* *"It's a problem with the structure of the query that you're using"*

3.2.1 Dominant Spike Occurrence Pattern

Some topics demonstrate a spike in addition to an amount of lower-level, relatively consistent communication activity. These topics are characterised by a higher-than-average occurrence duration, and a higher inter-quartile range of O_T. Here, the spike represents a need for further discussion of the topic than is typical due to specific circumstances or events in the project or the design, such as a design change or formal design meetings and delegation of actions. This is evidenced by looking in more detail at some of the email content associated with topics in Figure 5, see Table 8.

In contrast to single spike topics, those following dominant spike activity display a more consistent, albeit low, level of discussion. They include more commonly occurring topics and are of broader relevance across the project timeline, and hence are less likely to describe single events or bodies of work. It is plausible that, should issues or external events not have occurred, such topics would have shown similarities to background chatter across their lifespan.

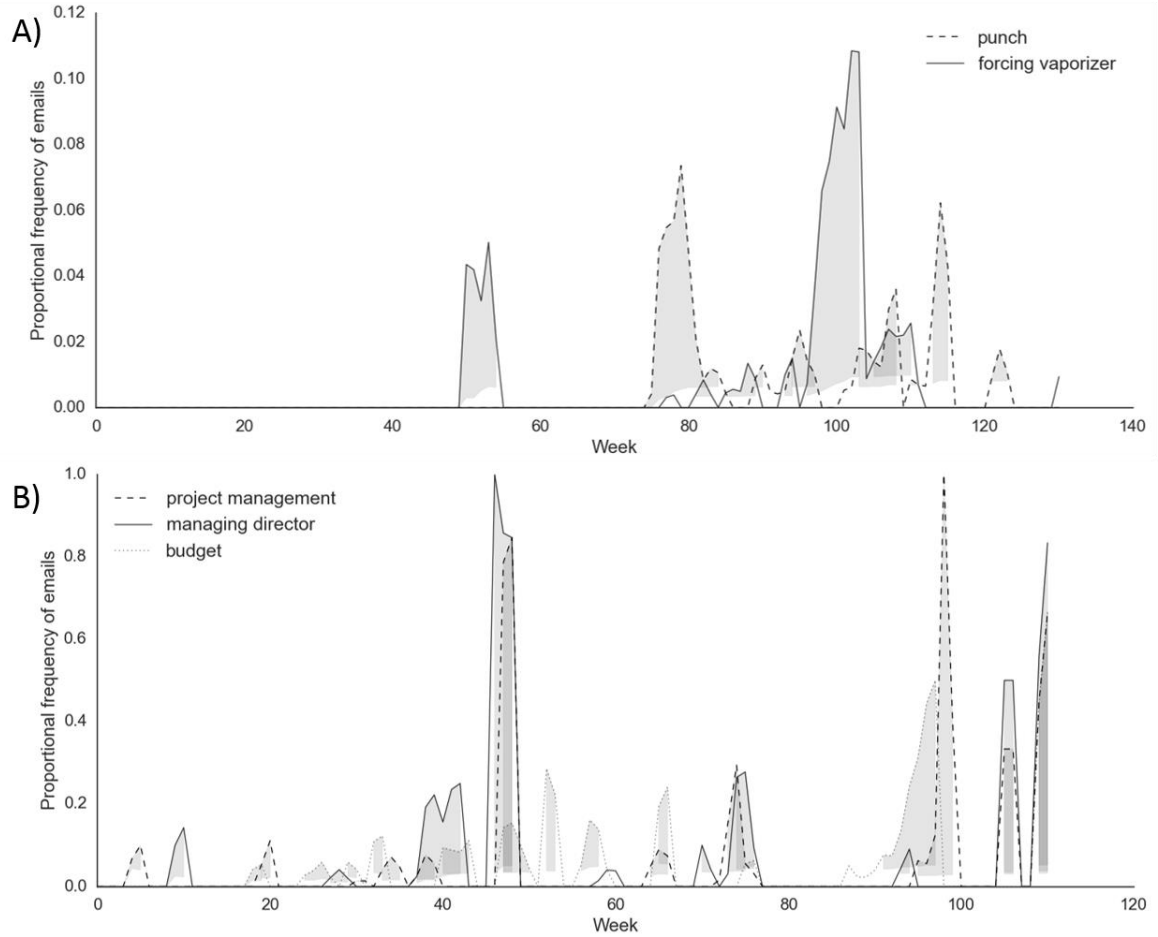


Figure 5: Dominant spike topics from Dataset A (above) and Dataset B (below). Shaded areas indicate a value of $O(T) > 1$.

Table 8: Context of topics demonstrating a dominant spike trace

Topic	Context
Project 1 – *Punch*	Development of an individual system. Spike occurs during and following first testing of the system, in which many results are reported, and actions formed, discussed, and delegated.
Project 1 – *Forcing Vaporizer*	Development of an individual sub-system. Initial spike (~ week 50) – discussion of a number of design changes that are required. Second spike (~ week 100) - discussion of testing procedures and requests for further information between geographically distributed parties.

3.3 Variable Occurrence Discussion Pattern

Many topics display a more variable occurrence, with bursts of high discussion and troughs of little to none. These have varying values of cumulative and relative occurrence through their life, varying duration, and are generally characterised by the appearance of a number of spikes punctuating periods of lower activity, see Figure 6.

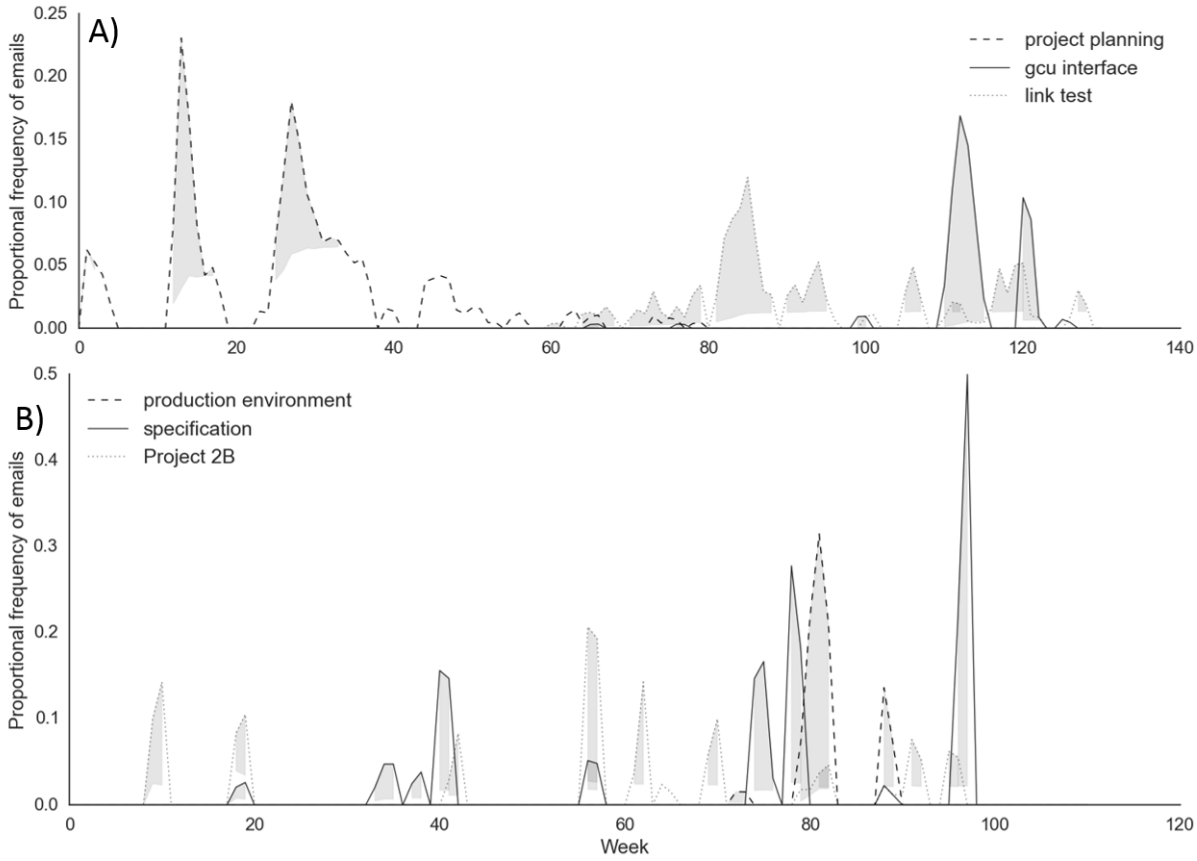


Figure 6: Variable occurrence topics from Dataset A (above), Dataset B (below). Shaded areas indicate a value of $O(T) > 1$.

Within some identified variable occurrence patterns there is evidence of periodicity, see *project planning* (Figure 6A). These regular periods of higher cumulative and relative occurrence may stem from regular events within the project – e.g. alignment with discussion of an upcoming meeting agenda and subsequent dissemination of outcomes.

Other spikes in variable occurrence patterns stem from periods in which work in that area was particularly focused. For the topic *link test* (Figure 6A), the spike at around week 85 occurs during the organisation and implementation of a major testing procedure operated by an external company. For the topic *specification* (Figure 6B), the spikes at around week 75 occur during the formation of a secondary specification for an additional interface.

These topics are characterised by inconsistency through their lifespans, with periods of higher and lower occurrence, multiple spikes punctuating periods of inactivity, and occasional periodic usage. In all cases such variance is determined by the context-specific requirements of the project, where project circumstances such as issues, higher levels of relevance, or logistical discussion have required the input of personnel in varying manners.

3.4 Summary Discussion

Topic occurrence within engineering communication shows a changeable landscape, with several dynamic patterns found in communication activity (see Table 9), and with significant overlap to those identified within other fields (Gruhl et al. 2004).

Table 9: Patterns in topic activity traces

Pattern	Features / comments
Background discussion	Consistent frequencies of usage throughout the topic life
Single spike discussion	High frequencies of usage for a brief period, surrounded by zero-usage periods
Dominant spike discussion	High frequencies of usage punctuating periods of lower-level, occasionally consistent usage
Variable discussion	Inconsistent usage, with higher and lower frequency periods, consistent and inconsistent periods, and elements of periodicity in use.

A number of topics align to background chatter, with a relatively consistent occurrence throughout their lifespan. These appear to represent discussion around core systems, concepts, and areas of the project that are pervasive throughout the process. These are the heart of the project – those areas that form the common threads around which the project occurs. Such topics may therefore provide potential for a manager to understand the core topics of their projects, as well as to monitor those that become more or less core over time.

In contrast to this consistency, many topics display higher occurrence bursts at certain points within their life. There are a number of patterns associated with these - a single spike amongst low-level activity, a single spike from no activity, periodic spikes, and high-variability activity. From the content of the emails under study, such spikes and variability appear to occur due to events or issues in the project, and often require broader discussion or effort from more personnel. These include standard project events such as discussion of meetings, logistics, and administration and events of more consequence, such as discussion around design changes, missing information, and project issues (See Sections 3.2, 3.3 for examples).

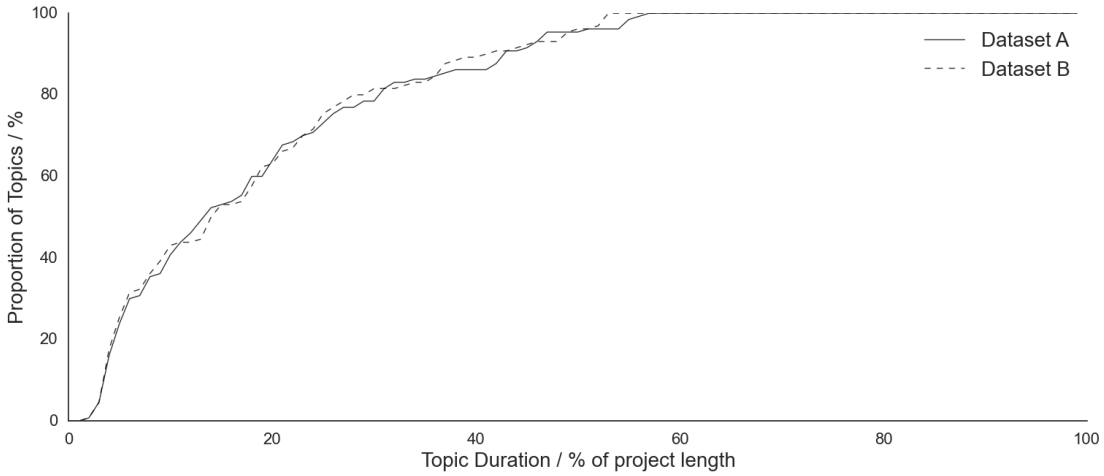
It should be noted however that as each topic is individual and distinct, there is a challenge in delineating specific boundaries between each pattern. For example, the height at which activity is said to create a spike, or variation required for background chatter to be classified as variable discussion are potentially fluid concepts, and in reality are likely fuzzy boundaries. Further investigation is needed to identify appropriate mathematical boundaries for classification, if feasible, and to ensure correct and useful information can be provided for the manager.

4 Topic Duration and Process Stage Localisation

While the previous section takes a topic-centric perspective, identifying patterns in the dynamic trace of each, this section studies the appearance of topics from a process-centric perspective, and forms a number of characterisations of process-dependent topic occurrence. Analysis focuses on the duration of occurrence of each topic, and the process stage or stages in which relative occurrence was high.

4.1 Topic Occurrence Duration through the Project

Figure 7 provides a summary of occurrence duration of individual topics through the lifespan of each project, hence representing the extent of time in which each topic had high relative occurrence, and was the subject of more focused discussion.



	Dataset A	Dataset B
Median	0.177	0.140
Upper Quartile	0.308	0.243
Lower Quartile	0.0692	0.0473

Figure 7: Chart of cumulative occurrence (above) and descriptive statistics (below) for topic occurrence duration. Median used as data is non-normal.

This analysis shows similarity between datasets; 50% of topics in each have an occurrence duration of less than 20% of the project length. Less than 20% exceeded 40% in duration, while none exceeded 60.0%. Periods of high occurrence for a typical topic can therefore be considered to be relatively short-lived, with most displaying low levels of activity through the majority of the project.

Table 10: Example topics of different occurrence durations

	Dataset A Topics	Dataset B Topics
UQ - Max	Company Acronym 1C, Project 1A, Company Acronym 1B3, Pressure	Company 2A1, Company Acronym 2B, Company 2B, Company 2A2, schedule, Company Acronym 2A
Median – UQ	pcrf, fuel gas, starboard, spare parts, Company Acronym 1B2, cargo handling, Company 1D	Itinerary, hosting, support specialist, Company 2D, Project 2A, workflow, resources director, software developer
LQ – Median	Cable interconnection, network switch, meeting minutes, lv switchboard, airfreight, voltage dip	Central services, customer implementation, requirements document, occupancy rules, test environment, certification process
Min – LQ	Electric drawings, characteristic curve, spray pipe, gas purging, interface test, battery voltage, ballast pump starter	Lowest fare, travel system, data setup, specific fare, generic access, real time flights, long haul flight, test data

4.2 Multiple-Stage Topic Occurrence

Across both datasets a large number of topics persist between and across process stages. Although some relation may be expected due to the distinct purposes of different stages and the tasks within (Pahl & Beitz 1984; Pugh 1990; Hales 1987), this demonstrates that many topics are not stage-bound, but are of relevance across the project.

A majority of topics are first raised in the initial two stages of the project, but are not resolved until later (85.8% Dataset A, 69.2% Dataset B emerging in first two stages; 8.76% Dataset A; 7.69% Dataset B disappearing in first two stages; see Table 11), perhaps reflecting planning work and suggesting that the majority of important topics within a project are raised early-on for future resolution. Further, very

few topics are first raised in the last project stage (0.302% Dataset A, 0.00% Dataset B, Table 11), suggesting that the work that occurs here is typically not new, but rather foreshadowed by that in earlier project stages.

Table 11: Topic occurrence in project stages

Dataset A	First appearance in stage (%)	Discussed in stage (%)	Dataset B	First appearance in stage (%)	Discussed in stage (%)
Specification	53.8	53.8	Stage 1	40.8	40.8
Manufacture	32.0	83.7	Stage 2	28.5	69.2
Sub-system Testing	10.3	87.3	Stage 3	8.46	70.0
Assembly	3.63	86.1	Stage 4	22.3	88.5
Testing	0.302	64.4	Stage 5	0.00	42.3

4.3 Stage Localisation of Topics

Many topics are only discussed within smaller segments of the project process. Taking the process stage boundaries in each dataset, the occurrence duration of each topic can be used to identify those for which discussion was primarily isolated within a single stage, was spread strongly amongst multiple stages, and those with more varied occurrence over the project life. A topic is defined as isolated to one or more stages by proportional occurrence in excess of that appearing in any combination of other stages. A topic is considered isolated in *n* stages if a) the proportion of its occurrence duration within each stage of the isolation group, P_i, is greater than the combined proportion of all stages outside of the group, P_o, by a factor, *f*, of (*n + 1*), and b) the topic is not part of an isolation group for any larger value of *n*:

$$P_i \geq \frac{P_o f}{n}; \quad where\ f = n + 1$$

The value of *f* chosen here allows maximum tolerance for lower-level discussion in other stages while ensuring a majority in isolated stages, thus allowing interpretation of topic / stage relationships. In industry application, a value that represents isolation should be determined in the specific context. Figure 8 shows topics from Dataset A that are isolated within one (*commercial proposal* - specification stage; *Company 1A2* - sub-system testing) and two stages (*Company Acronym 1C2* - assembly and testing).

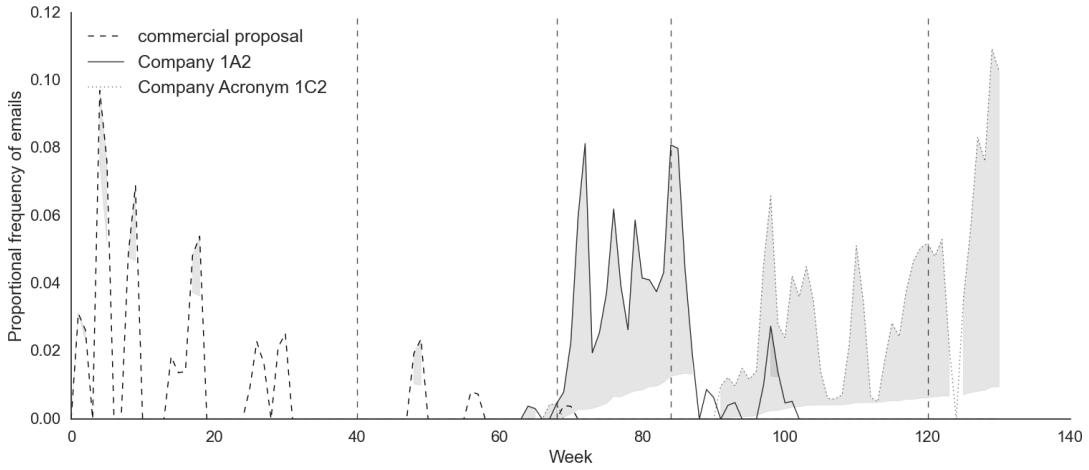


Figure 8: Isolated topics within Dataset A

In each dataset the majority of topics are not isolated to any combination of stages, with a variable relative occurrence across the process (64.7% Dataset A, 52.3% Dataset B; Table 12). This supports the

finding that topic discussion is not typically stage-specific, but rather is of relevance throughout the project process. With that said, even when non-isolated to project stages, the majority of high relative occurrence in each dataset occurs during central stages (Manufacture - Assembly, Dataset A; Stage 2 – Stage 4, Dataset B; Table 12), indicating that these stages contain the highest levels of focused work and a wider breadth of work areas. That very few topics are equally important in many stages (3/4/5 stage focused topics; Table 12) suggests that *background chatter* topics, despite generally consistent in appearance, will show variation that is not coincident with stage boundaries. The peak seen for *Company Acronym 1C2* in Figure 8 is likely due to the decrease in email transmission at the end of the project, thus inflating the relative proportion of topics sent from the company that remained active.

Table 12: Proportion of isolated topics

Dataset A	**Number of stages in which topic is primarily focused**					
	1 Stage	2 Stages	3 Stages	4 Stages	5 Stages	Non-Isolated
Proportion isolated (%)	18.4	10.3	2.72	2.42	1.51	64.7
Stages in which topic discussion is focused						
Specification (%)	3.28	10.3	11.1	3.13	20.0	9.25
Manufacture (%)	39.3	29.4	29.6	25.0	20.0	28.1
Sub-system Testing (%)	36.1	22.1	29.6	25.0	20.0	24.0
Assembly (%)	4.92	25.0	18.5	25.0	20.0	25.0
Testing (%)	16.4	13.2	11.1	21.9	20.0	13.7

Dataset B	Number of Stages					
	1 Stage	2 Stages	3 Stages	4 Stages	5 Stages	Non-Isolated
Proportion isolated (%)	37.7	8.46	0.00	1.54	0.00	52.3
Stages in which topic discussion is focused						
Stage 1 (%)	0.00	0.00	--	0.00	--	1.92
Stage 2 (%)	10.2	36.4	--	25.0	--	22.1
Stage 3 (%)	6.12	36.4	--	25.0	--	21.2
Stage 4 (%)	83.7	27.3	--	25.0	--	48.1
Stage 5 (%)	0.00	0.00	--	25.0	--	6.73

When showing some traits of isolation, and hence more likely to be of specific relevance to the stages in which they appear, most topics belong to either one stage (18.4% Dataset A, 37.7% Dataset B; Table 12) or two (10.3% Dataset A, 8.46% Dataset B; Table 12), and typically belong to manufacture/testing (Dataset A) or stages 2/3 (Dataset B) – the central period of the project process. It is therefore plausible that in this period the work performed aligns with the specific purposes of the relevant process stages. Further, it implies that the topics within these stages are more often short-lived, following a *spike* pattern, and hence are likely to describe lower level systems.

Within Dataset B, a predominance of isolated and emerging topics can be found in stage 4 which, when studying the textual content of the emails, typically refer to terminology surrounding a complex software testing regime spread across multiple companies and systems. Their formation and discussion is a timely response to the needs of the project, indicating that larger bodies of work may initiate a new period of topic creation. That such topics were not raised in earlier stages could be a result of the more agile process methodology of the software development project, a lack of necessity for specific technical language before work commenced, or potentially a deficiency in appropriate planning of the testing procedure. Indeed, when looking at the content of the emails sent surrounding these topics, there is evidence of significant querying, organisation between distributed parties, and smaller issue resolution (see Table 13).

Table 13: Context of emails including *fare query* topic, isolated in stage 4; Dataset B

Context
In response to creation of an account to interface with a testing database: *"I cannot commit to a date when it will be ready, as it is not our team who carries out these requests, but based on past experience, I would think it should be ready early next week."*
In response to a series of clarification requests: *"We've been through your comments and have added our replies below. I think the most salient point is that this is a tour operator reservation system, […], and so the message flow is a little different from what one might imagine."*
In response to querying of an error message: *"I will send you an email tomorrow morning with the status of the resolution and we can decide from there how best to proceed with the certification."* *"The definitive fix will be put into place as soon as possible (so there might still be some slight instabilities until that is done) - however, this should not block your certification testing."*
From the company CEO on completion of the test process: *"Well done guys on getting this through the certification process. It'll be a major step forward in the process of getting […] live."*

4.4 Summary Discussion

This section has presented traits of topic occurrence in relation to project process stages, through the measurement of the periods in which relative occurrence was high. While a large amount of variation can be found between topics, the findings do suggest a typical pattern and set of characteristics to the topic discussion within an engineering project, which have potential to be of direct use to a project manager in their work.

Topics with a higher occurrence duration (above ~20%, Section 4.1) show a tendency towards high level and core project systems and concepts. This aligns with individual topic patterns identified, specifically the longer-lived and core subjects apparent in background chatter.

There is little evidence for a strong stage-specificity in the occurrence of topic discussion when following stages delineated by project members. Despite the variation in purpose of different stages and the types of task that each require, communication activity associated with specific topics appears to transcend stage boundaries, indicating that the tasks to which many individual topics are subject encompass those common in multiple process stages. Interestingly, and a subject for future work, that topics do not conform to stage boundaries may also act as commentary on the quality and reality of boundaries set by personnel within a project. The characteristics of topic activity presented here allow the building of general trends of topic activity for engineering projects, see Table 14.

Table 14: Characteristics of topic communication activity through the project process

Stage	Characteristic
Early	The majority of topics that will be discussed throughout the project emerge. Very few topics are resolved.
	Topics typically become *background chatter*, or follow either *dominant spike* or *variable occurrence* patterns.
	Topics that emerge are typically not low-level systems, concepts, or components.
Central	Many topics are short-lived, likely in the form of *spikes*.
	The widest breadth of topics are discussed.
	Low-level topics will typically occur and be discussed.
Late	Few new topics emerge.
	A majority of topics that are present emerged during earlier stages, and are here resolved.
	Topics do not typically follow a *spike* pattern, and are typically not low-level.

5 General Discussion

This work has identified a number of patterns in the dynamics of topic activity through a project timeline, and typical characteristics of topic discussion that occur with particular reference to process stage boundaries, see Table 15. While there is a clear need for further investigation of different project scenarios, confirmation, and extension of characteristics that have been identified, the analysis methods and findings of this work demonstrate both feasibility and potential value of the approach employed.

Engineering projects prove to be highly variable in the patterns that activity follows within, with different forms occurring dependent on level of generality, system / component level, closeness to the core of the project, specific project events such as deliverables or issues, and process stage. By identification of patterns and characteristics of topic activity, the results of this work begin to provide a grounding on which a manager's understanding may develop, and a characterisation of engineering process activity on which future research may build.

Table 15: Patterns of topic activity traces and characteristics of high-focus discussion

Pattern	**Features / comments**
Background discussion	Consistent frequencies of usage throughout the topic life
Single spike discussion	High frequencies of usage for a brief period, surrounded by zero-usage periods
Dominant spike discussion	High frequencies of usage punctuating periods of lower-level, occasionally consistent usage
Variable discussion	Inconsistent usage, with higher and lower frequency periods, consistent and inconsistent periods, and elements of periodicity in use.
Characteristic	
High-level concepts and systems are highly discussed for longer periods of a project, while low-level systems, components, and concepts are short-lived.	
Topic activity is typically not directly linked to project stages, or stage-specific activity types.	
The majority of topics emerge during early stages, and persist beyond.	
Central project stages contain highest levels of discussion concerning the broadest range of topics.	
Low-level topics are typically discussed during central project stages.	
Topics discussed in later stages typically emerged earlier in the project process	
Individual events within a project can greatly increase focus on individual topic areas.	

In line with the stated benefits of this work such patterns and characterisations give potential to benefit a project manager in two streams, in the archival analysis of past projects for purposes of future planning, and in the real-time monitoring of in-progress work.

By post-hoc analysis and characterisation, managers are able to increase their historic understanding and evidence-base future planning and decisions. For example, identification of topics of differing pattern may indicate the extent of those that are core at different stages, hence aiding resource planning. Where a topic has previously become core in certain stages, a manager may plan to increase resource in future projects. Where certain topics are identified as having followed spike or dominant spike patterns, which indicate specific project events and potentially issues, a manager may focus additional investigative effort to identify root cause and form lessons learned for future projects. For example, the multi-spike pattern of *"forcing vaporizer"*, Fig. 5, demonstrated a rise-fall-rise pattern in activity with certain stages requiring higher focus and broader input. Knowledge of the cause and periodicity of this pattern in the specific or general case may aid in schedule planning, resource distribution, and provide foci for managerial attention in future projects. Similarly, location of individual topics in process stages, and quantity of such topics in each, may allow a manager to organise personnel and resource such that demands in future projects can be better met – i.e. ensuring that personnel with aligned capabilities are

available at the appropriate time. In all cases, the detailed analytic approaches employed denote the possibility for detailed understanding to be generated.

Particularly with reference to topic patterns and the activity a manager may expect to appear, the analysis also provides scope for real-time monitoring and management of engineering projects. The project events indicated by certain patterns give scope for rapid highlighting of potentially problematic work areas. For example, the situations arisen for spike pattern topics in Section 3.2 display evidence of difficulty and may require rapid management. That the algorithms developed in this analysis may automatically highlight the appearance of such a pattern allows potential for immediate managerial intervention. Although non-problematic, the appearance of other patterns may allow a manager to focus their attention to increase project effectiveness. For example, the high activity indicated by a dominant spikes (see Section 3.2.1) may alert a manager that a specific project event is underway, and monitoring of its progress; and periodicity in variable occurrence patterns may allow automatic monitoring of activity levels in line with expected periodic trends. Scope also exists for the manager to play a more dominant role, in which the patterns and stage-based characteristics of each topic are presented and compared to typical values. Should the manager, through their detailed contextual knowledge, identify incongruences with their or historical expectation they are able to focus their attention to ensure activity is proceeding as planned. For example, should a topic that is typically core background chatter start to display a variable occurrence or spike occurrence, should certain topics or patterns appear outside of their typical stage, or should the appearance of certain patterns break with their typical stage location. In each case, the analysis employed provides scope to focus attention towards the "non-normal" in the specific project context, thereby potentially drawing the manager towards areas in which their input is particularly important and providing evidence upon which they may base their decision making processes.

Key to implementation of this real-time approach, and a current limitation of this work for both real-time and historical analysis, is need for a manual parsing stage to identify specific topics from the output of the *tf-cidf* algorithm. For appropriate implementation in industry, this process must be performed by personnel embedded in the project context. In this work, to enable broad identification of patterns, effort was made towards inclusiveness. This issue may be mitigated in the first case by generation of a project-specific topic list by embedded personnel, in which experience and historical cases inform those topics that should be monitored. Particular interest may then arise from those topics detected as important by the algorithm that are outside of the generated list. Second, the topic list must be regularly updated by periodic implementation of the *tf-cidf* algorithm and manual parsing at a rate such that topics are not missed. While not time-consuming, the value of analysis and additional associated workload must be weighed against the learnings and benefits generated in each individual case. Finally, the *tf-cidf* algorithm employed produces a breadth of viable topics and mitigates issues in topic evolution and change, but lacks subtlety and semantic context; as a result, multiple topics identified may in reality belong to a higher-level group of words with similar meaning or implication, thereby conflating the proportion of each pattern identified. Exploration of semantic topic modelling methods is viewed as a valuable future development to understand the relation between lower-level topics, and the patterns evident in semantically linked topic groups throughout the project timeline.

Context-specific understanding is key in generating value. Every engineering project presents a distinct situation, will have varying factors that influence their performance, and will demonstrate different patterns in the activity that occurs within them. While some consistency has been found in the topic activity of both Datasets A and B variation should always be expected, characteristics may hold more or less true in other cases, and new common trends may emerge. This underlines the vital role of the project manager in data interpretation. Where benefit derives from a deep understanding of the project

and performance within a specific context, there is a need for a manager with deep knowledge to judge. The inherent variation between project scenarios is a significant challenge to interpretation of performance by algorithmic means - while data analysis can be quantifiable and general, interpretation should be guided by theoretical considerations, with the sense-making and holistic understanding of projects generated by reflective interpretation of results. The approach and analyses presented here are therefore for the purpose of support of this subjective interpretation; allowing a manager to increase understanding of their own project while minimising the effort they must commit to do so, encouraging judgement of performance in each specific case, and thereby allowing the manager to focus their effort on improvement and intervention, rather than investigation.

Several avenues for future work exist, both to increase confidence in results and to extend understanding gained. The work presented here constitutes early-stage research, and requires further exploration in other datasets and in collaboration with acting project managers for detail and validation. Several specific developments would be highly beneficial. First, in-depth analysis of the dynamics of a single project, enabling detailed understanding of features within topic traces and mathematical delineation of patterns found through, for example, signal processing and feature recognition techniques. Such detailed patterns and features within should be corroborated by interview and observational analysis of the project in-progress. Second, further study and comparison in additional engineering contexts of varying scale, global distribution, sector, complexity, etc., to determine extent of consistency in patterns and characteristics identified. Third, extraction and study of further communication data, both physical and digital, in order to extend the dataset, including for example social and short-text communications, digital work such as reports and wikis, and inter-personal communications such as telephone and face-to-face conversation. This may be particularly relevant in the discussion of sensitive matters that may be conducted offline. Further, while all employees were required to include emails in the extracted mailboxes by company mandate, this cannot be guaranteed. Higher quantities of data would increase granularity and confidence in results, ensuring consistency in the proportions of each pattern found. Finally, scope exists for extension with additional analysis, such as semantic topic detection (Dumais et al. 1988; Dumais 2004), sentiment analysis (Pang & Lee 2008; Thelwall et al. 2010), and correlation of analysis against actual project performance. These may prove invaluable resources for broader contextual and situational understanding.

6 Conclusions

This work has performed a detailed analysis of the discursive characteristics of a broad range of topics, extracted directly from industry engineering projects within the marine engineering and software engineering fields. By analysing the cumulative occurrence, relative occurrence, and duration of occurrence of a number of topics extracted from the email communications of workers within each project, this work proposes a number of patterns and characteristics of engineering communication activity. The analytics developed and patterns and characterisations identified have potential to benefit engineering project management though study of historic data for lessons learned and future planning, as well as for real-time engineering project monitoring.

Using the dynamic activity of each topic, extracted through the cumulative and relative levels at which each topic was discussed, distinct patterns of activity were found with differing implications for the role of the topic in the wider project context. These include identification of background chatter within a project and the core themes that run through it, of low-level areas of work that are short-lived and specific in nature, and of identification of patterns that may prove signatory of issues requiring manager attention. By monitoring the periods of time over which each topic was discussed highly, a number of characteristics of activity in relation to the project process and its individual stages were formed. These

clarify the emergence of new topics, the type of pattern to be expected throughout different process stages, and the extent to which topic activity can be considered stage-specific throughout an engineering project process.

Through the ability to monitor and analyse real engineering projects, and compare data to the patterns and characteristics presented here, a manager is able to increase their understanding of the projects over which they have responsibility. This is particularly important given the contextual variation rife across project scenarios, and subsequent challenge in finding general rules for high performance that can be applied in a general case. The analyses presented here act as a support method, providing the means for provision of detailed project information to a manager, enabling their planning and decision-making processes, and hence allowing the streamlining of management and intervention to enhance and support project, process, and product improvement.

Acknowledgements

The work reported in this paper has been undertaken as part of the Language of Collaborative Manufacturing (www.locm.org.uk) Project at the University of Bath & University of Bristol, which is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant reference EP/K014196/2. It was supported by the Croatian Science Foundation (www.hrzz.hr) through the MInMED project (www.minmed.org). Data are available at the University of Bristol data repository, data.bris, at https://doi.org/10.5523/bris.3oj79e4nco7qk2u14xg719ezga

7 References

Ahmed, S., Wallace, K.M. & Blessing, L.T., 2003. Understanding the differences between how novice and experienced designers approach design tasks. *Research in engineering design*, 14(1), pp.1–11.

Allan, J., 2002. *Topic Detection and Tracking*, New York: Springer Science+Business Media.

AlSumait, L., Barbará, D. & Domeniconi, C., 2008. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp.3–12.

Aral, S., Brynjolfsson, E. & Van Alstyne, M., 2007. Productivity Effects of Information Diffusion in Networks.

Atkinson, R., 1999. Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria. *International Journal of Project Management*, 17(6), pp.337–342. Available at: http://www.sciencedirect.com/science/article/pii/S0263786398000696.

Banker, R.D., Bardhan, I. & Asdemir, O., 2006. Understanding the Impact of Collaboration Software on Product Design and Development. *Information Systems Research*, 17(4), pp.352–373.

Brants, T., Chen, F. & Tsochantaridis, I., 2002. Topic-based document segmentation with probabilistic latent semantic analysis. *Proceedings of the eleventh international conference on Information and knowledge management CIKM 02*, p.211. Available at: http://portal.acm.org/citation.cfm?doid=584792.584829.

Byron, K., 2008. Carrying too heavy a load? The Communication and Miscommunication of Emotion by Email. *The Academy of Management Review*, 33(2), pp.309–327.

Cantamessa, M., Montagna, F. & Neirotti, P., 2010. Understanding the organizational impact of PLM systems. *International Journal of Operations & Production Managemen*, 32(2), pp.191–215.

Carley, K.M., 2003. Dynamic Network Analysis. *Dynamic social network modeling and analysis Workshop summary and papers*, pp.133–145. Available at: http://www.chronicdisease.org/files/public/2009Institute_NA_Track_Carley_2003_dynamicnetwork.pdf.

Cataldi, M. et al., 2010. Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In *THe 10th International Workshop on Multimedia Data mining*. Available at: http://dl.acm.org/citation.cfm?id=1814245.1814249.

Cha, M., Mislove, A. & Gummadi, K.P., 2009. A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. *Proceedings of the 18th International Conference on World Wide Web*, pp.721–730. Available at: http://doi.acm.org/10.1145/1526709.1526806.

Chapman, C. & Ward, S., 1996. *Project risk management: processes, techniques and insights*, Chichester, UK: John Wiley & Sons, Ltd.

Chen, H., Chiang, R.H.L. & Storey, V.C., 2012. Business Intelligence and analytics: From Big Data to Big Impact. *MIS Quarterly Quarterly*, 36(4), pp.1165–1188. Available at: http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=c72752a6-fd0c-4184-ad0b-39ae0c9c16d8@sessionmgr4003&vid=1&hid=4209.

Collins, A. & Baccarini, D., 2004. Project Success - A Survey. *Journal of Construction Research*, 5(2), pp.211–231. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=14874051&site=ehost-live.

Cooke-Davies, T., 2002. The "real" success factors on projects. *International Journal of Project Management*, 20(3), pp.185–190. Available at: http://www.sciencedirect.com/science/article/pii/S0263786301000679.

Coursey, K. & Mihalcea, R., 2009. Topic identification using Wikipedia graph centrality. In *Proceedings of NAACL HLT 2009*. Boulder, Colorado, pp. 117–120. Available at: http://portal.acm.org/citation.cfm?doid=1620853.1620887.

Cross, N. & Cross, A.C., 1998. Expertise in engineering design. *Research in engineering design*, 10(3), pp.141–149.

Dewar, R.D. & Dutton, J.E., 1986. The Adoption of Radical and Incremental Innovations: An Empirical Analysis. *Management Science*, 32(11), pp.1422–1433. Available at: http://www.jstor.org/stable/2631501.

Dorst, K. & Cross, N., 2001. Creativity in the design process: co-evolution of problem-solution. *Design Studies*, 22(5), pp.425–437. Available at: http://www.sciencedirect.com/science/article/B6V2K-43BXNJY-3/2/ec88c05634af799b89a41081220036b2.

Dumais, S.T., 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1). Available at: http://www.scholarpedia.org/article/Latent_semantic_analysis.

Dumais, S.T. et al., 1988. Using Latent Semantic Analysis to Improve Access to Textual Information. *ACM Conference on Human Factors in Computing Systems, CHI '88*, (October 2014), pp.281–285.

Eagle, N. & Pentland, A., 2006. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), pp.255–268.

Earl, C., Eckert, C. & Clarkson, J., 2005. Design Change and Complexity. In *2nd Workshop on Complexity in Design and Engineering*. Glasgow, Scotland.

Engwall, M., 2002. No project is an island: linking projects to history and context. *Research Policy*, 32(2003), pp.789–808.

Fernandez, D.J. & Fernandez, J.D., 2008. Agile Project Management —Agilism versus Traditional Approaches. *Journal of Computer Information Systems*, 49(2), pp.10–17. Available at: http://www.tandfonline.com/doi/abs/10.1080/08874417.2009.11646044.

Floricel, S. & Miller, R., 2001. Strategizing for anticipated risks and turbulence in large-scale engineering projects. *International Journal of Project Management*, 19(8), pp.445–455.

Gero, J.S. & Tang, H.H., 2001. The differences between retrospective and concurrent protocols in revealing the process-oriented aspects of the design process. *Design Studies*, 22(3), pp.283–295.

Gruhl, D. et al., 2004. Information diffusion through blogspace. *ACM SIGKDD Explorations Newsletter*, 6, pp.43–52.

Guille, A. & Hacid, H., 2012. A predictive model for the temporal dynamics of information diffusion in online social networks. *Proceedings of the 21st international conference …*. Available at: http://dl.acm.org/citation.cfm?id=2188254.

Gupta, A. et al., 2009. Use of collaborative technologies and knowledge sharing in co-located and distributed teams: Towards the 24-h knowledge factory. *Journal of Strategic Information Systems*, 18(3), pp.147–161.

Hales, C., 1987. *Analysis of the Engineering Design Process in an Industrial Context*. Cambridge: University of Cambridge.

Highsmith, J. & Cockburn, A., 2001. Agile software development: The business of innovation. *Computer*, 34(9), pp.120–122.

von Hippel, E., 1998. Economics of Product Development by Users: The Impact of "Sticky" Local Information. *Management Science*, 44(5), pp.629–644.

Holme, P. & Saramaki, J., 2012. Temporal networks. *Physics Reports*, 519(3), pp.97–125.

Iribarren, J.L. & Moro, E., 2009. Impact of human activity patterns on the dynamics of information diffusion. *Physical Review Letters*, 103(July), pp.8–11.

Johansen, L. et al., 2007. Email Communities of Interest. *Ceas*. Available at: http://www.cse.psu.edu/~butler/pubs/ceas07.pdf.

Jones, S. et al., 2014. Finding Zelig in Text : A Measure for Normalizing Linguistic Accommodation. In *COLING 2014: 25th International Conference on Computational Linguistics*. Dublin, Ireland.

Jones, S. et al., 2015. Subject Lines As Sensors : Co-Word Analysis Of Email To Support The Management Of Collaborative Engineering Work. In *ICED'15: International Conference on Engineering Design*.

Kuhn, A., Ducasse, S. & Girba, T., 2007. Semantic clustering: Identifying topics in source code. *Information and software technology*, 49(3), pp.230–243.

Labrinidis, A. & Jagadish, H. V., 2012. Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, 5(12), pp.2032–2033. Available at: http://dl.acm.org/citation.cfm?id=2367572%5Cnhttp://dl.acm.org/citation.cfm?doid=2367502.2367572.

Landaeta, R.E., 2008. Evaluating Benefits and Challenges of Knowledge Transfer Across Projects. *Engineering Management Journal*, 20(1), pp.29–38. Available at: http://0-search.ebscohost.com.library.vu.edu.au/login.aspx?direct=true&db=bth&AN=32526091&site=bsi-live.

Lavagnon, A.I., 2009. Project success as a topic in project management journals. *Project Management Journal*, 40(4), pp.6–19.

Matsubara, Y. et al., 2012. Rise and fall patterns of information diffusion: model and implications. *the 18th ACM SIGKDD international conference*, pp.6–14. Available at: http://dl.acm.org/citation.cfm?doid=2339530.2339537.

May, R.M. & Lloyd, A.L., 2001. Infection dynamics on scale-free networks. *Physical Review E*, 64, pp.66112–66114. Available at: http://link.aps.org/abstract/PRE/v64/e066112.

McAlpine, H., Hicks, B.J. & Culley, S.J., 2009. Comparing the information content of formal and informal design documents: lessons for more complete design records. *ICED09: International conference on engineering design*.

Munson, S.A., Kervin, K. & Robert, L.P., 2014. Monitoring email to indicate project team performance and mutual attraction. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pp.542–549. Available at: http://dl.acm.org/citation.cfm?id=2531602.2531628.

Myers, S.A., Zhu, C. & Leskovec, J., 2012. Information diffusion and external influence in networks. In *KDD 2012*. Beijing, China, p. 33. Available at: http://dl.acm.org/citation.cfm?doid=2339530.2339540%5Cnhttp://www.scopus.com/inward/record.url?eid=2-s2.0-84866021278&partnerID=tZOtx3y1.

Pahl, G. & Beitz, W., 1984. *Engineering Design: A Systematic Approach*, London: Springer.

Pang, B. & Lee, L., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), pp.1–135. Available at: http://www.nowpublishers.com/article/Details/INR-001.

Parraguez, P., Eppinger, S.D. & Maier, A.M., 2015. Information Flow Through Stages of Complex Engineering Design Projects: A Dynamic Network Analysis Approach. *IEEE Transactions on Engineering Management*, 62(4), pp.604–617.

Pastor-Satorras, R. & Vespignani, A., 2001. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14), pp.3200–3203.

Patrashkova-Volzdoska, R.R. et al., 2003. Examining a curvilinear relationship between communication frequency and team performance in cross-functional project teams. *IEEE Transactions on Engineering Management*, 50(3), pp.262–269.

Pinto, J.K. & Slevin, D.P., 1987. Critical factors in successful project implementation. *Engineering Management, IEEE Transactions on*, (1), pp.22–27.

Pugh, S., 1990. *Total Design: integrated methods for successful product engineering*, Harlow: Prentice Hall.

Robertson, S., 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), pp.503–520.

Robinson, M.A., 2010. An empirical analysis of engineers' information behaviours. *Journal of the American Society for Information Science and Technology*, 61(4), pp.640–658.

Robinson, M.A., 2012. How design engineers spend their time: Job content and task satisfaction. *Design Studies*, 33(4), pp.391–425. Available at: http://dx.doi.org/10.1016/j.destud.2012.03.002.

Rodan, S. & Galunic, C., 2004. More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal*, 25(6), pp.541–562.

Romero, D.M., Meeder, B. & Kleinberg, J., 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *Proceedings of the 20th international conference on World wide web*, pp.695–704. Available at: http://dl.acm.org/citation.cfm?id=1963503.

Schmidt, R. et al., 2001. Identifying software project risks: an international Delphi study. *Journal of management information systems*, 17(4), pp.5–36.

Škec, S., Cash, P. & Štorga, M., 2017. A dynamic approach to real-time performance measurement in design projects. *Journal of Engineering Design*, 4828(March), pp.1–32. Available at: https://www.tandfonline.com/doi/full/10.1080/09544828.2017.1303665.

Snider, C. et al., 2016. Determining Work Focus, Common Language, and Issues in Engineering Projects Through Topic Persistence. In *DESIGN 2016: International Conference on Engineering Design*. Dubrovnik, Croatia.

Snider, C., Mcalpine, H., et al., 2015. It's Not Personal: Can Logbooks provide insights into engineering

projects? In *ICED'15: International Conference on Engineering Design*. Milan, Italy.
Snider, C., Gopsill, J.A., et al., 2015. Understanding Engineering Projects: An Integrated Vehicle Health Management Approach to Engineering Project Monitoring. In *ICED15: International Conference on Engineering Design*. Milan, Italy.
Spärck Jones, K., 1972. A Statistical Interpretation of Term Specificity and its Retrieval. *Journal of Documentation*, 28(1), pp.11–21. Available at: http://www.emeraldinsight.com/doi/abs/10.1108/eb026526.
Tang, J. et al., 2010. Analysing Information Flows and Key Mediators through Temporal Centrality Metrics Categories and Subject Descriptors. *Proceedings of the 3rd Workshop on Social Network Systems*, (Figure 1), p.3.
Thelwall, M. et al., 2010. Sentiment Strength Detection in Short Informal Text. *The American Society for Informational science and technology*, 61(12), pp.2544–2558. Available at: http://www.scit.wlv.ac.uk/~cm1993/papers/SentiStrengthPreprint.pdf.
Toor, S.-R. & Ogunlana, S.O., 2010. Beyond the "iron triangle": Stakeholder perception of key performance indicators (KPIs) for large-scale public sector development projects. *International Journal of Project Management*, 28(3), pp.228–236.
Uflacker, M. & Zeier, A., 2011. A semantic network approach to analyzing virtual team interactions in the early stages of conceptual design. *Future Generation Computer Systems*, 27(1), pp.88–99. Available at: http://dx.doi.org/10.1016/j.future.2010.05.006.
Wallace, K.M. & Ahmed, S., 2003. How Engineering Designers Obtain Information. In U. Lindemann, ed. *Human behaviour in design. Individuals, teams, tools*. Munich: Springer-verlag, pp. 184–194.
Wang, T., Rao, J. & Hu, Q., 2014. Supervised word sense disambiguation using semantic diffusion kernel. *Engineering Applications of Artificial Intelligence*, 27, pp.167–174.
Wasiak, J. et al., 2010. Managing by E-Mail : What E-mail Can Do for Engineering Project Management. *IEEE Transactions on Engineering management*, pp.1–12.
Wasiak, J. et al., 2010. Understanding engineering email: The development of a taxonomy for identifying and classifying engineering work. *Research in engineering design*, 21(1), pp.43–64.
Watson, J., 2012. Keynote address at the University of Bath.
Wu, F. et al., 2004. Information flow in social groups. *Physica A*, (1), pp.327–335. Available at: http://www.hpl.hp.com/research/idl/papers/flow/flow.pdf.
Xia, W. & Lee, G., 2004. Grasping the complexity of IS development projects. *Communications of the ACM*, 47, pp.68–74.